

ABSTRACT

This thesis focuses on four main research themes namely shot boundary detection, fast frame alignment, activity-driven video summarisation, and highlights based video annotation and retrieval. A number of novel algorithms have been proposed to address these issues, which can be highlighted as follows.

Firstly, accurate and robust shot boundary detection is achieved through modelling of cuts into sub-categories and appearance based modelling of several gradual transitions, along with some novel features extracted from compressed video. Secondly, fast and robust frame alignment is achieved via the proposed subspace phase correlation (SPC) and an improved sub-pixel strategy. The SPC is proved to be insensitive to zero-mean-noise, and its gradient-based extension is even robust to non-zero-mean noise and can be used to deal with non-overlapped regions for robust image registration. Thirdly, hierarchical modelling of rush videos using formal language techniques is proposed, which can guide the modelling and removal of several kinds of junk frames as well as adaptive clustering of retakes. With an extracted activity level measurement, shot and sub-shot are detected for content-adaptive video summarisation. Fourthly, highlights based video annotation and retrieval is achieved, in which statistical modelling of skin pixel colours, knowledge-based shot detection, and improved determination of camera motion patterns are employed.

Within these proposed techniques, one important principle is to integrate various kinds of feature evidence and to incorporate prior knowledge in modelling the given problems. High-level hierarchical representation is extracted from the original linear structure for effective management and content-based retrieval of video data. As most of the work is implemented in the compressed domain, one additional benefit is the achieved high efficiency, which will be useful for many online applications.

DEDICATION

吾生也有涯，而知也无涯

*Our life is limited, but the knowledge
to explore is boundless.*

—庄子 (from Chuang-tzu, 369BC–286BC)

*This thesis is dedicated to my wife Yujing,
my son Tianqi and my Parents.*

*Without their constant love, encouragement and support,
this work would have no chance been possible.*

ACKNOWLEDGEMENTS

First of all, I would like to thank Prof. J. Jiang and Dr. S. S. Ipson for their valuable guidance and ever support throughout my PhD study. The relatively loose research environment has brought me great freedom in exploring several different topics and facing the relevant challenges. Numerous discussions on all types of research problems with them have led me to this success.

I would also like to thank all my colleagues and friends in Bradford, especially Miss Juan Chen, Ms Chunmei Qing, and Dr. Ying Weng for the support and many useful discussions with them. Special thanks are owing to Miss Rona Wilson for her kind support.

Though having been dedicated before, I have to again express my grateful appreciation to my family, especially to my wife Yujing as well as my son, my parents and others. It is their strong support and encouragement that has helped me to complete this study.

I am very grateful for Prof. H. Hu at Essex University and Dr. D. Rigas to be my external and internal examiners. Their comments and suggestions have improved the quality of this thesis and made it more fluent and readable.

Finally, I wish to acknowledge the financial support from the EU IST FP6 Project “LIVE” (Contract No. IST-4-027312) and also thank the TRECVID community for the wonderful platform they have provided for me to develop and evaluate the proposed techniques.

TABLE OF CONTENTS

ABSTRACT	i
DEDICATION	ii
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	x
LIST OF TABLES	xiii
1. INTRODUCTION	1
1.1 Background.....	1
1.2 Problems.....	3
1.3 Research Objectives.....	6
1.4 Research Methodologies.....	6
1.5 Thesis Contributions.....	8
1.6 Thesis organisation.....	10
2. LITERATURE REVIEW	13
2.1 Introduction.....	13
2.2 Video Segmentation Techniques.....	14

2.3	Frame Alignment Techniques.....	18
2.4	Video Summarisation Techniques.....	20
2.4.1	Previous Work in Video Summarisation.....	21
2.4.2	Summarisation of Rush Videos in TRECVID.....	22
2.5	Video Annotation and Retrieval.....	24
2.5.1	Motion Pattern Determination.....	25
2.5.2	Skin and Human Object Detection.....	26
2.6	Summary.....	30
3.	MODEL-BASED SHOT BOUNDARY DETECTION	33
3.1	Introduction.....	33
3.2	Compressed Domain Feature Extraction.....	34
3.2.1	Feature Extraction.....	34
3.2.2	Feature Analysis.....	37
3.3	Pre-filtering of Cuts and Feature Selection.....	39
3.3.1	Pre-filtering of Cuts.....	39
3.3.2	Feature Selection.....	41
3.4	Modelling and Detecting Abrupt Shot Changes.....	43
3.4.1	Modelling.....	43
3.4.2	Decision Rules for the Five Categories of Cuts.....	48
3.4.3	Validation of Detected Cuts	51
3.4.4	Determining Parameters.....	54

3.5 Determining Gradual Transitions and Fusion.....	56
3.5.1 Determining Combined Cuts.....	56
3.5.2 Determining Other Gradual Transitions.....	57
3.5.3 Fusion of Detected Results.....	59
3.6 Results and Discussions.....	60
3.6.1 Data Preparation.....	60
3.6.2 Overall Performance and Evaluation.....	61
3.6.3 Performance Analysis in Detail.....	65
3.7 Summary.....	72
4. FAST AND ROBUST FRAME ALIGNMENT	73
4.1 Introduction.....	73
4.2 Phase Correlation Method.....	74
4.3 Subspace Phase Correlation.....	76
4.3.1 Subspace Phase Correlation.....	76
4.3.2 Peak Height Analysis.....	78
4.3.3 Robustness Analysis.....	78
4.4 Implementation and Sub-pixel Accuracy.....	80
4.4.1 Dealing with Non-overlapped Regions.....	80
4.4.2 Subpixel Accuracy.....	83
4.4.3 Further Improving Robustness.....	85
4.5 Results and Discussions.....	86

4.5.1	Synthetic Data.....	87
4.5.2	Real MRI Data.....	89
4.5.3	Robustness Analysis.....	92
4.5.4	Moving Sequence Data.....	94
4.5.5	Computational Complexity.....	97
4.6	Summary.....	97
5.	ACTIVITY-DRIVEN VIDEO SUMMARISATION	99
5.1	Introduction.....	99
5.2	Modelling and Video Structuring.....	100
5.2.1	Modelling.....	100
5.2.2	Shot Detection and Activity-Level Determination.....	102
5.2.3	V-Units Determination.....	104
5.3	Filtering Junk Frames.....	106
5.3.1	Filtering H-Cut Frames.....	107
5.3.2	Filtering Clapboards from s_clip Frames.....	108
5.3.3	Filtering e_clip Frames.....	109
5.4	Determining Retakes.....	110
5.4.1	Similarity of Selected Key-Frames.....	110
5.4.2	Adaptive Shot Clustering.....	112
5.5	Generating Video Summaries.....	114
5.6	Results and Discussions.....	118

5.6.1	Data Set and Evaluation Criteria.....	118
5.6.2	Overall Results.....	121
5.6.3	Intermediate Results.....	126
5.6.4	Error Analysis.....	130
5.7	Summary.....	131
6.	HIGHLIGHTS-BASED ANNOTATION AND RETRIEVAL	133
6.1	Introduction.....	133
6.2	Feature Extraction and Video Segmentation.....	136
6.2.1	Feature Extraction from MPEG Videos.....	136
6.2.2	Knowledge-Based Shot Change Detection.....	138
6.3	Skin Detection.....	140
6.3.1	Modelling Skin and Non-skin Colours in Compressed Domain.....	141
6.3.2	Bayesian Classification.....	143
6.3.3	Optimal Thresholding.....	144
6.3.4	Post-processing.....	147
6.3.5	Experimental Results on Skin Detection.....	148
6.4	Annotation and Retrieval of Video Highlights.....	151
6.4.1	Determining Camera Motion Patterns.....	152
6.4.2	Highlights-Based Annotation and Retrieval.....	154
6.5	Results and Discussions.....	154
6.5.1	Results in Video Segmentation.....	156

6.5.2	Results in Skin and Human Object Detection.....	157
6.5.3	Results in Determination of Motion Patterns.....	158
6.5.4	Results in Video Highlights Annotation and Retrieval.....	159
6.6	Summary.....	162
7.	CONCLUSIONS AND FUTURE WORK	163
7.1	Introduction.....	163
7.2	Main Contributions.....	164
7.3	Future Work.....	167
APPENDIX 1.	Author's Contributions.....	171
REFERENCES	174

LIST OF FIGURES

3.1. Example of one cut in four consecutive frames with the original frame images (top) and their corresponding DC images (bottom)	37
3.2 Three DC-differencing images of each two consecutive DC images in Fig. 3.1 and their corresponding binary masks after adaptive thresholding (Column a-b). Results in (c) are those by median filtering of (a) and their binary masks are given in (d).....	38
3.3. Typical samples of five categories of cuts and one exception (the last row) in consecutive four frames of the video sequences.....	45
3.4. Workflow of the system for shot boundary detection.....	47
3.5. Determining η , where $a_c(p_c) = a_{\bar{c}}(p_{\bar{c}}) = 0.01$ and p'_0 is the middle point of p_c and $p_{\bar{c}}$	55
3.6. Precision-recall curves to show effectiveness of selected parameters where ρ , p_0 and t_μ for cut detection and e_0 for gradual transition detection	66
3.7. Examples of missing detected cuts due to similar intensity (a), strong motion (b) and undefined in the five categories of cut (c)	68
3.8. Four examples of false detected cuts of apparent visual changes in frames.....	68
3.9. Six arguable cuts in ground truth where four cuts are listed in the top two rows and two non-cuts at the bottom.....	69
3.10. Examples of defined gradual transitions (top row) and undefined arguable gradual transitions (bottom two rows).....	70
4.1. Overlapped and non-overlapped areas in two images.....	81
4.2. Definition of five sub-images in each test image.....	85
4.3. Two test images used to generate subpixel shifts.....	87
4.4. Three examples of test images: (a) original MRI image; (b) and (c) are two noisy versions of (a) with additive Gaussian noise.....	90

4.5. Mean square errors (y-axis) vs. Gaussian variance (x-axis)	93
4.6. Average height of the most dominant peak (y-axis) vs. Gaussian variance (x-axis)	93
4.7. Original two images from “flower garden” sequence at frame #20 and #21 and their raw difference (top row) and three motion-compensated results (bottom), estimated by, from left to right, using Foroosh [149], the proposed 2-D and subspace phase correlation, respectively.....	95
4.8. Performance in terms of SNR (in dBs) vs frame number using global motion compensation for the “flower garden” sequence.....	96
5.1. Formal description of the hierarchical model	101
5.2. Explaining the concepts of “vUnit”, “valid frames” and “active frames” using plotted curve of activity level (y-axis) vs. frames (x-axis), where representative frames are also shown which are categorized into normal frame and three kinds of junk frames (H-cut, s_clip and e_clip)	106
5.3. Examples of two junk frames from H-cut category (top) and their associated luminance histograms (bottom)	107
5.4. Three frames to show the process of moving clapboard in and out of the scene (a) and change of energy during this process (b)	108
5.5. Generated frames for video summarization with embedded texts and artificial one-frame dissolve	117
5.6. Examples of first eight cuts detected from MRS336853.mpg sequence with their start/end frames and associated cut likelihoods	126
5.7. List of key frames extracted for each V-unit of the first shot containing five retakes	129
5.8. List of final summarization results of the five retakes in Figure 5.6 and Table 5.3	129
6.1. An overall system diagram for video highlights extraction for automatic annotation and content-based retrieval	134
6.2. One macroblock in 4:2:0 chrominance format contains four luminance subblocks and two chrominance subblocks (a) and each subblock has 8*8 pixels (b).....	142
6.3. Histograms of logarithm likelihood ratio of skin and non-skin colours.....	145

6.4. Curves of A_s and $A_{\bar{s}}$ against logarithm likelihood ratio indicates potential missing detection rate and false alarm rate.....	146
6.5. Examples of four test frames (a) and their associated masks of skin (b), non-skin (c) and don't care pixels (d).....	148
6.6. Four results of detected skin from images in Fig. 6.5. (a) and (b) are results from proposed method using thresholds 12.2 and 49.25, respectively; (c) and (d) are results from Sigal etc[112] using their static and dynamic models, respectively.....	150
6.7. Plots of F_1 measures on the detection of cuts, GTs and all shots in the eight test sequences (a) to (h).....	156
6.8. Three groups of skin detection results (a-c), where the top row shows original image frames and corresponding skin results are shown at the bottom. Extracted human objects are labelled within red boxes.....	158
6.9. Main query interface with list of main characters automatically extracted via human object detection.....	160
6.10. Sketches to illustrate various motion patterns: the first row is for pure pan/tilt; the second row is for mixed shift; and the last refers to zooming in/out and static cases.....	161
6.11. Retrieval performance measure by precision-recall curves in terms of objects, events and combined cases with objects under certain events.....	162

LIST OF TABLES

3.1. Performance comparison using AdaBoost based cross validation on the data from TRECVID in 2006 and 2005.....	43
3.2. Percentage of cuts in different categories.....	61
3.3. Summary of test videos in TRECVID 2005-2007.....	61
3.4. Comparing results in terms of performance in detection of all transitions and cuts at TRECVID 2007, and only the best one from each participant is listed in decreasing F1 measure order. The “SysID” of the proposed system is with the prefix “AIS” which are highlighted with gray cell background.....	62
3.5. Comparing results in terms of performance in detection of gradual transitions at TRECVID 2007, and only the best one from each participant is listed in decreasing F1 measure order. The “SysID” of the proposed system is with the prefix “AIS” which are highlighted with gray cell background.....	63
3.6. Results of shot detection from the data in TRECVID 2006 and 2005.....	66
3.7. Effect of post-processing in cut detection.....	67
3.8. Comparisons of complexity and speed.....	71
4.1. Table of results for shifts of the images in Fig. 4.3 using linear interpolation.....	89
4.2. Pair-wise registration results of the five MRI images.....	91
4.3. Pair-wise registration results of the MRI images using subspace phase correlation without the use of local gradient	91
5.1. Descriptions of nine criteria used for evaluation.....	120
5.2. Typical results from TRECVID’08 on BBC rush summarization in decreasing IN score order.....	124
5.3. Obtained shot candidates corresponding to the detected cuts in Figure 5.6.....	127
6.1. Performance comparisons of the proposed approach and Sigal etc. in [112].....	151
6.2. Detail information about the test sequences.....	155
6.3. Performance in terms of precision and recall measures for shot detection.....	156
6.4. Precision and recall measures for camera events detection.....	159